ED 451 210 TM 032 437

AUTHOR	Brooks, B. Meade
TITLE	How To Handle Discrete Dependent Variables in the Univariate Case: A Primer on Logistic Regression.
PUB DATE	2001-02-02
NOTE	19p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (24th, New Orleans, LA, February 1-3, 2001).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Predictor Variables; *Regression (Statistics)
IDENTIFIERS	*Dependent Variables; *Logistic Regression

ABSTRACT

This paper presents an overview of logistic regression and illustrates the method with the data transformations that are conducted. It also discusses the interpretation of logistic regression results. To make the discussion more concrete, an analysis of a data set is presented in which logistic regression is used to predict the likelihood of a college student's withdrawing or failing a course. Logistic regression is a well-suited analysis technique when a dichotomous dependent variable is involved because the logit transformation allows for direct linear comparisons of the effect different indicators have on the outcome. Logistic regression also aids in defining quantitatively the combination of predictors that leads to varying degrees or probabilities of the outcome variable. Applying the approach to identifying high-risk students in educational settings is especially practical. (SLD)

ED 451 210

How to Handle Discrete Dependent Variables in the Univariate Case: A Primer on Logistic Regression

B. Meade Brooks

University of North Texas

Brookhaven Community College

Presented at:

24th Annual Meeting

SERA (Southwest Educational Research Association)

New Orleans, Louisiana

February 2, 2001

TM032437

Abstract

Many variables in social science research occur naturally in continuous form. For example, attitudes, intelligence, and personality are often measured at an interval level of scale. Of course, not all variables occur in continuous form (e.g., gender, pass/fail grading, etc.). When nominal variables naturally occur, traditional analysis of variance-type methods are certainly warranted. However, when nominal data happen to be the dependent variable, both ANOVA and multiple regression techniques are insufficient. The purpose of this paper is to: a) present an overview of logistic regression, b) illustrate the method along with the data transformations that are conducted, and c) provide discussion concerning how to interpret logistic regression results. To make the discussion more concrete, analysis of a data set will be presented in which logistic regression is used to predict the likelihood of a college student withdrawing or failing a course.

Introduction

Several analytical options are available for examining discrete dependent variables, including discriminant analysis, logit multiway frequency analysis, and logistical regression. Of

these methods, Tabachnick and Fidell (1996) noted that logistic regression is the most flexible because,

> Unlike discriminant function analysis, logistic regression has no assumptions about the distributions of the predictor variables; in logistic regression, the predictors do not have to be normally distributed, linearly related, or of equal variance within each group. Unlike multiway frequency analysis, the predictors do not need to be discrete; the predictors can be any mix of continuous, discrete and dichotomous variables. Unlike multiple regression analysis, which also has distributional requirements for predictors, logistic regression cannot produce negative probabilities. (p. 575)

Furthermore, logistic regression can handle dependent variables with more than two outcomes that may or may not have an inherent ordinal ranking. It should be noted, however, that logistic regression is essentially a univariate analysis and is limited to a single dependent variable (Huck, 2000). By contrast, discriminant analysis is multivariate in nature (Klecka, 1980).

Logistic regression approaches have been popular in some fields (e.g., health/medical sciences) due to the nature of the research questions often asked in these settings (Tabachnick & Fidell, 1996). The discrete outcome in logistic regression is often disease/no disease. For example, can risk of heart disease be predicted from blood pressure and age? Logistic regression is especially useful when the distribution of responses on the dependent variable is expected to be nonlinear with one or more of the independent variables. For example, the probability of heart disease may be little affected (say 1%) by a 20-point

difference among people with low blood pressure (e.g., 105 vs. 125) but may change quite a bit (say 5%) with an equivalent difference among people with high blood pressure (e.g., 190 vs. 210).

One reason for logistic regression's popularity lies with its ability to explain outcomes in terms of an odds ratio (Huck, 2000). The odds ratio can be intuitively understood (e.g. "the participants were about 5 times more likely to have. . .") by most applied researchers. However, the mathematical foundation of the ratio and the equations underlying logistic regression are a bit more complex. Because the outcome variable is discrete in nature, special data transformations, called logits, are required.

Theory

The basic premise behind multiple regression analysis (MRA) is that a continuous outcome variable is, in theory, a linear combination of a set of predictors and error. Thus, for an outcome variable, Y, and a set of n predictor variables, $X_1, \ldots, X_n$, the MRA model is of the form:

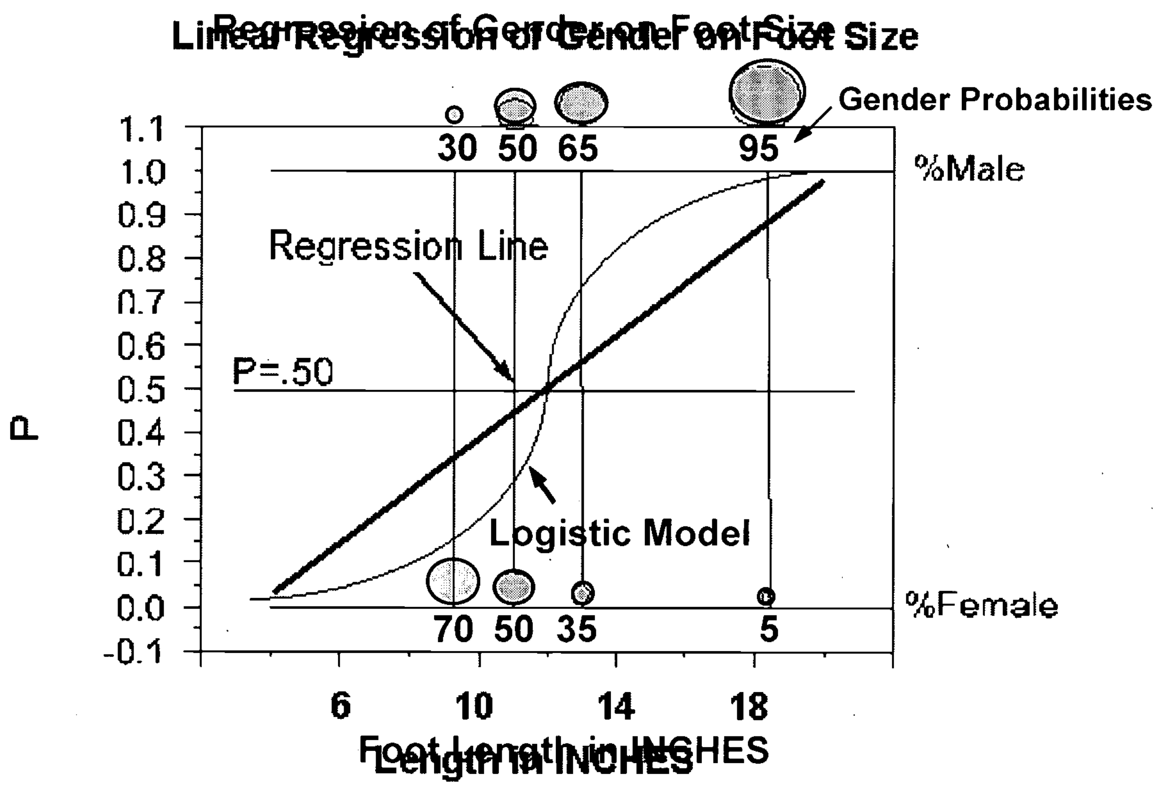
$$Y = \alpha + \beta X_1 + \beta X_2 + \ldots + \beta_n X_n + \varepsilon = \alpha + \sum_{J=1}^{n} \beta_j X_j + \varepsilon$$

where α is the Y-intercept (i.e., the expected value of Y when all X's are set to 0), β_v is a multiple (partial) regression coefficient (i.e., the expected change in Y per unit change in X_n assuming all other X's are held constant) and ε is the error of prediction. If error is omitted, the resulting model represents the expected, or predicted, value of Y. We can interpret the MRA model as follows: each observed score, Y, is made up of an expected, or predictable, component that is a function of the predictor variables $X_1,...,X_p$, and an error, or unpredictable component that represents error of measurement (i.e., unreliability) and/or error in the selection of the model.

Logistic regression is a variation of ordinary regression, useful when the observed outcome is restricted to two values, which usually represent the occurrence or non-occurrence of some outcome event, (usually coded as 1 or 0, respectively). It produces a formula that predicts the probability of the occurrence as a function of the independent variables. Just like linear regression, logistic regression gives each regressor a coefficient β that measures the regressor's independent contribution to variations in the dependent variable. But there are technical problems with dependent variables that can only take values of 0 and 1.

Suppose we want to predict whether someone is male or female (DV, M=1, F=0) using his or her foot size in inches (IV). We could plot the relations between the two variables as we customarily do in regression. The plot might look something like Figure 1. The Y-axis is P, which indicates the proportion of 1's at any given value of height

FIGURE 1



Notice that none of the observations actually fall on the linear regression line. They all fall on zero or one. The regression model will also allow estimates below 0 and above 1. The predicted values will become greater than one and less than zero

if you move far enough on the X-axis. Such values are theoretically inadmissible. Another problem lies in that one of the assumptions of regression is that the variance of Y is constant across values of X (homoscedasticity). This cannot be the case with a binary variable, because the variance is PQ. When 50 percent of the people are 1's, then the variance is .25, its maximum value. As we move to more extreme values, the variance decreases. When P=.10, the variance is .1*.9 = .09, so as P approaches 1 or zero, the variance approaches zero.

A hint at the solution lies in observing that a logarithmic curve best "fits" the data. The logistic transformation of *p*, also called taking the logit of *p*, is the log (to base *e*) of the odds or likelihood ratio that the dependent variable is 1. In symbols it is defined as:

$$\text{logit}(p)=\log(p/(1-p))$$

Whereas *p* can only range from 0 to 1, logit(*p*) ranges from negative infinity to positive infinity. The logit scale is symmetrical around the logit of 0.5 (which is zero). The logit transformation spreads out the differences between extreme probabilities; the differences of logits between gender likelihoods of .95 and .99 is much bigger than that between .5 and .7. The success of logistic regression is based on the

characteristic that the logit transformation changes the non-linear probability scale into a linear "logit" scale.

It follows that logistic regression involves fitting to the data an equation of the form:

$$\mathrm{logit}(p) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots$$

Although logistic regression finds a "best fitting" equation just as linear regression does, the principles on which it does so are rather different. Instead of using a least-squared deviations criterion for the best fit, it uses a maximum likelihood method, which maximizes the probability of getting the observed results given the fitted regression coefficients. A consequence of this is that the goodness of fit and overall significance statistics used in logistic regression are different from those used in linear regression.

The logistic regression model is identical to the multiple regression model except that the log-odds in favor of Y = 1 replaces the expected value of Y. There is a relatively simple exponential transformation for converting log-odds back to probability:

$$P = \frac{1}{1 + \exp[-(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots)]}$$

The *odds* of an event is defined as the probability of the outcome event occurring divided by the probability of the event not occurring. The odds ratio for a predictor tells the relative amount by which the odds of the outcome increase (O.R. greater than 1.0) or decrease (O.R. less than 1.0) when the value of the predictor value is increased by 1.0 units.

Methodology

Logistic regression forms a predictor variable that is a linear combination of the explanatory variable. The values of this predictor variable are then transformed into probabilities by a logistic function. To illustrate this process, analysis is presented of a data set consisting of various predictors characteristic of students at Brookhaven College in the Dallas Community College District. The goal of the logistic regression model is to predict whether or not a student will fail or drop a course (as opposed to receiving an A, B, C, or D) based on basic characteristics of the course and student. These course/student indicators include gender, residency status, ethnicity, TASP testing status, credit hours of course being taken, type of course taken, date of course registration, number of students in the class, and number of weeks in the course. Of particular

interest is the ability to identify and predict the success of "high-risk" students. High-risk students are defined as being more likely to be unsuccessful in a course for various reasons. Because the logistic regression model can calculate individual probabilities of dropping/failing a class for each student, it provides a quantitative way to identify students most likely to be unsuccessful in class, i.e. high risk. It is important to note that logistic regression analysis does not suggest that an indicator (such as gender) "causes" high or low probability of event occurrence. Logistic regression, like all regression analyses, is based on correlations or relationships between variables, which in many cases are indirect. A relationship does not imply cause.

Results

PC SAS was used to run the logistic regression analysis of the data that consisted of over 18,000 courses taken Fall of 1999. Of the model's predicted probabilities, 67.8% where concordant with the model, and 31.8% where discordant. This gives an initial impression of a moderate fit to the data. Tables 1 and 2 show other output relevant to determining the success of the model. The Wald Chi-Square statistic indicates that the model overall was statistically significant (Table 1), meaning it did better than if someone where to simply guess with

50-50 odds of a correct prediction. The large data set and moderate fit also led to most of the indicators having a statistically significant impact on the model (Table 2 - Chi-Square). Table 2 also gives the individual logit coefficients (Estimate) and intercept for the logistic equation.

TABLE 1 - The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1778.3700	29	<.0001
Score	1629.6605	29	<.0001
Wald	1474.1096	29	<.0001

TABLE 2 - Analysis of Maximum Likelihood Estimates

Parameter	DF	β Coeff	Error	Chi-Square	Pr>ChiSq
Intercept	1	-2.9090	0.4525	41.3196	<.0001
Reg_Diff	1	-0.0168	0.000981	293.2318	<.0001
NO_WEEKS	1	0.0455	0.00755	36.2428	<.0001
NO_STUDENT	1	-0.00543	0.000963	31.8188	<.0001
AGE	1	-0.0128	0.00209	37.8408	<.0001
CRED 1	1	-0.3759	0.0830	20.5340	<.0001
CRED 2	1	-0.1400	0.1808	0.5997	0.4387
CRED 3	1	-0.1660	0.0704	5.5553	0.0184
CRED 4	1	0.2571	0.0756	11.5710	0.0007
TYPE 1	1	0.3373	0.0362	86.5930	<.0001
TYPE 2	1	0.0192	0.0470	0.1663	0.6834
RESIDENCY 1	1	0.3111	0.0446	48.6582	<.0001
RESIDENCY 2	1	0.2083	0.0489	18.1273	<.0001
RESIDENCY 3	1	0.2387	0.1018	5.4997	0.0190
ETHNICITY 1	1	0.1508	0.0557	7.3236	0.0068
ETHNICITY 2	1	0.1660	0.0668	6.1734	0.0130
ETHNICITY 3	1	0.0434	0.0610	0.5076	0.4762
ETHNICITY 4	1	-0.1316	0.0696	3.5795	0.0585
ETHNICITY 5	1	0.6133	0.1611	14.4868	0.0001
ETHNICITY 6	1	-0.3644	0.1013	12.9468	0.0003

TABLE 2 continued…

Parameter		DF	β Coeff	Error	Chi-Square	Pr>ChiSq
GENDER	F	1	-0.1151	0.0163	50.0788	<.0001
MP	0	1	0.3489	0.0204	291.9072	<.0001
MPB	0	1	0.4803	0.2000	5.7672	0.0163
MRC	0	1	0.4151	0.0513	65.3768	<.0001
RF	0	1	-0.4760	0.0307	240.2258	<.0001
RFB	0	1	0.5827	0.2675	4.7436	0.0294
RP	0	1	-0.3799	0.0247	236.3726	<.0001
RPB	0	1	-0.2911	0.0644	20.4301	<.0001
WFB	0	1	0.7360	0.2763	7.0977	0.0077
WRC	0	1	0.2852	0.0450	40.2318	<.0001

The analysis now deviates from typical checks for logistic regression model success. Of primary interest in this study is the ability to accurately predict high-risk students. To best evaluate this model, predictions for dropping or failing a course where made for Fall 2000 students based on the logit equation derived from the Fall 1999 data. Again, a similar number (67.4%) of the predictions where accurate, or "matched" the model. But most of the "bad" predictions occurred around probabilities of 50%, which is also the probability region in which most students fell. Figures 2 and 3 illustrate. Notice that at the extreme probabilities (less than 10% = successful students, more than 70% = high-risk students), the number of correct predictions is much higher than between, say, 30% and 60%. Table 3 illustrates the accuracy of the model by showing

FIGURE 2

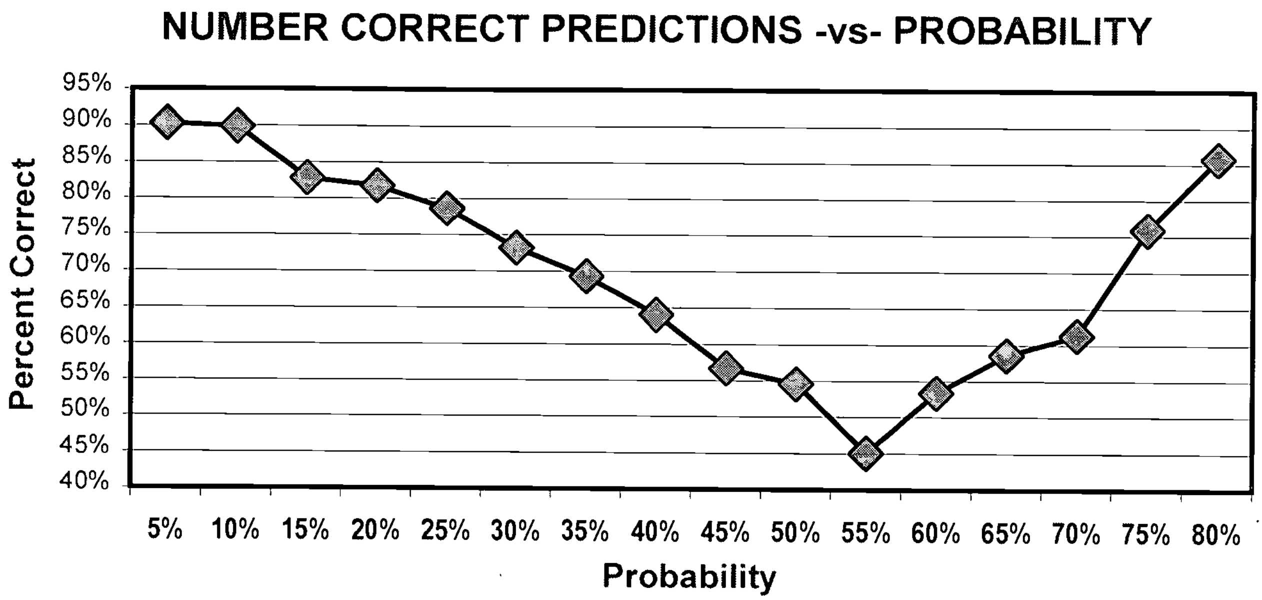


FIGURE 3

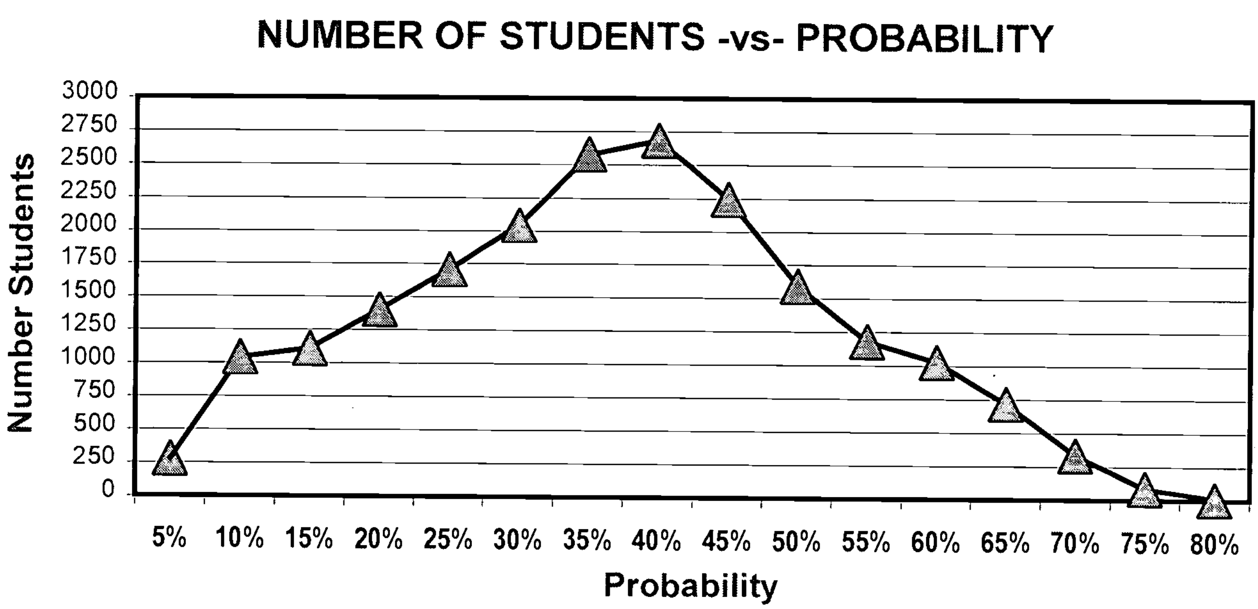


Table 3 - Accuracy of Logistic Model for High/Low Probabilities

"AT-RISK" STUDENTS

LOGIT W or F	Prob W or F	Predict W or F	Actual W or F
1.25	77.80%	1	1
1.20	76.80%	1	0
1.19	76.70%	1	1
1.14	75.70%	1	1
1.14	75.70%	1	1
1.10	75.00%	1	1
1.10	75.00%	1	1
1.10	74.90%	1	1
1.09	74.90%	1	1
1.09	74.80%	1	1
1.09	74.80%	1	1
1.09	74.80%	1	1
1.08	74.70%	1	1
1.08	74.70%	1	1
1.05	74.10%	1	1
1.05	74.10%	1	0
1.05	74.10%	1	1
1.05	74.00%	1	1
1.02	73.60%	1	1
1.01	73.30%	1	1
1.01	73.20%	1	1
1.01	73.20%	1	0
0.99	73.00%	1	1
0.99	73.00%	1	1
0.99	72.90%	1	1
0.99	72.90%	1	0
0.98	72.70%	1	1
0.96	72.40%	1	1
0.96	72.30%	1	1
0.96	72.30%	1	1
0.96	72.20%	1	1
0.95	72.20%	1	1
0.95	72.20%	1	1
0.95	72.20%	1	1
0.95	72.10%	1	1
0.95	72.10%	1	1
0.94	71.90%	1	0
0.93	71.70%	1	1

"SUCCESSFUL" STUDENTS

LOGIT W or F	Prob W or F	Predict W or F	Actual W or F
-3.64	2.60%	0	1
-3.64	2.50%	0	0
-3.65	2.50%	0	0
-3.65	2.50%	0	0
-3.68	2.50%	0	0
-3.69	2.40%	0	0
-3.69	2.40%	0	0
-3.70	2.40%	0	1
-3.70	2.40%	0	0
-3.72	2.40%	0	0
-3.73	2.30%	0	0
-3.73	2.30%	0	0
-3.77	2.20%	0	0
-3.78	2.20%	0	0
-3.78	2.20%	0	0
-3.78	2.20%	0	0
-3.78	2.20%	0	0
-3.79	2.20%	0	0
-3.80	2.20%	0	1
-3.81	2.20%	0	0
-3.82	2.20%	0	0
-3.82	2.20%	0	0
-3.88	2.00%	0	0
-3.91	2.00%	0	0
-3.91	2.00%	0	0
-3.94	1.90%	0	1
-3.96	1.90%	0	0
-3.97	1.90%	0	1
-4.11	1.60%	0	0
-4.64	1.00%	0	0
-4.64	1.00%	0	0
-4.64	1.00%	0	0
-4.65	1.00%	0	0
-4.66	0.90%	0	0
-10.24	0.00%	0	0
-10.66	0.00%	0	0
-10.66	0.00%	0	0
-11.13	0.00%	0	0

predictions for the actual data for the highest and lowest 38 cases. Although the prediction model does only a moderate job overall, for those high-risk and successful students, it does very well. Incorporating better predictors may greatly improve even these results.

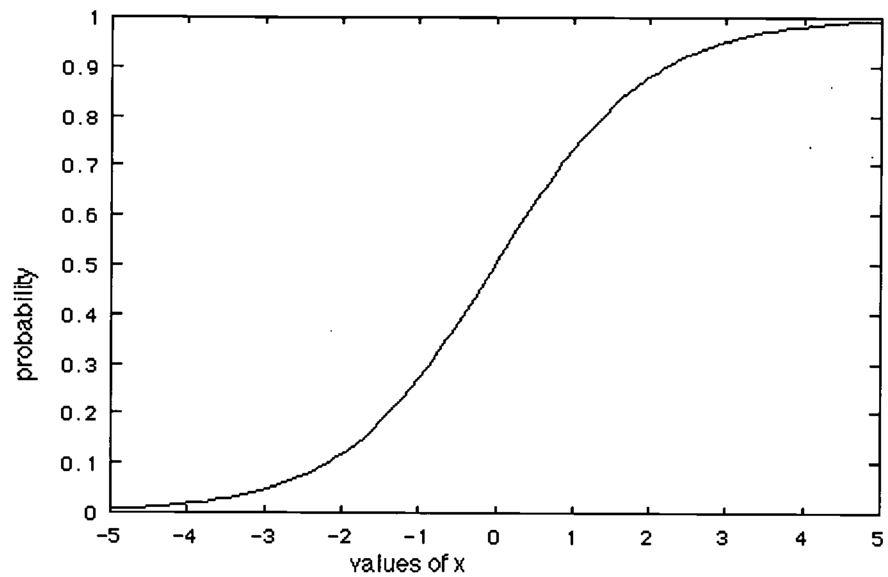
The model is further useful in that it allows the comparison of various configurations of indicators by observing the effect they have on the logit of the probability of failing or withdrawing. Table 4 shows the effect of changing various indicators, while holding the others constant to achieve a high-risk status. Because the logit scale is a transformation of probabilities to a linear scale, the ratio of two logits indicates how many more times likely the event will occur. For example, the ratio of the male to female logits indicates that males are 1.3 times as likely to fail or withdraw as females.

A constant increase in logit(p) has a reasonably straightforward interpretation. It corresponds to a constant *multiplication* (by $\exp(\beta)$) of the odds that the dependent variable takes the value 1 rather than 0. This leads to a convenient way of representing the results of logistic regression by a plot (Figure 4) showing the odds change produced by unit changes in different independent variables.

TABLE 4 - Logit Ratios

		LOGITS	LIKELIHOOD RATIOS	
			Compared to Next	Compared to Lowest
Gender				
	Male	**1.01**	1.3	
	Female	**0.78**		
Credit Hours for Course				
	Cred 4	**0.78**	2.3	5.4
	Cred 3	**0.34**	2.3	
	Cred 1	**0.15**		
Number Days until End of Registration				
	3	**1.21**	1.6	
	30	**0.76**		
Number Weeks in Course				
	32	**1.94**	1.6	
	16	**1.21**		
Course Type				
	1 - General Academic	**1.53**	1.3	1.8
	2	**1.21**	1.4	
	4	**0.83**		
Residency				
	In District	**1.53**	1.1	3.3
	Out of State	**1.46**	1.0	3.2
	Out of District	**1.43**	3.1	
	Out of Country	**0.46**		
# Students in Course				
	5	**1.56**	1.1	
	20	**1.47**		
Ethnicity				
	American Indian/Alaskan	**1.53**	1.4	3.5
	African-American	**1.08**	1.0	2.5
	White	**1.07**	1.1	2.4
	Hispanic	**0.96**	1.2	2.2
	Asian/Pacific Islander	**0.78**	1.4	1.8
	Non-Resident Alien/Foreign National	**0.55**	1.3	
	Unknown	**0.44**		
Age				
	22	**1.07**	1.2	1.7
	35	**0.90**	1.4	
	55	**0.64**		
MP				
	No	**1.07**	2.9	
	Yes	**0.37**		

FIGURE 4 - Relationship of Probability to β's



Conclusion

Logistic regression is a particularly well-suited analysis technique when a dichotomous dependent variable is involved. The logit transformation allows for direct linear comparisons of the effect different indicators have on the outcome. Logistic regression also aids in defining quantitatively what combination of predictors leads to varying degrees, or probabilities, of the outcome variable. The application to identifying high-risk students in educational settings is especially practical. Refinement of the logistic model presented will hopefully lead to improved accuracy in predicting students at high-risk of academic failure, with successful interventions to follow.

References

Klecka, W. R. (1980). Discriminant analysis. Newbury Park, CA: Sage.

Tabachnick, B. G., & Fidell, L. S. (1996). Using multivariate statistics (3rd ed.). New York: HarperCollins.

TM032437

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: How to Handle Discrete Dependent Variables in the Univariate Case: A Primer on Logistic Regression	
Author(s): Brooks, B. Meade	
Corporate Source: University of North Texas	Publication Date: Feb. 2001

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☒	Level 2A ☐	Level 2B ☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, →please

Signature: B. Meade Brooks	Printed Name/Position/Title: B. Meade Brooks - Res. Assoc.	
Organization/Address: UNT - Dept. of Technology & Cognition PO Box 311337, Denton, TX 76203-1337	Telephone: 940-565-2093	FAX: —
	E-Mail Address: bmb2114@dcccd.edu	Date: 2/12/01

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse: University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com